Mineração de Dados Complexos
Curso de Aperfeiçoamento
INF-0617 - Big Data

Processamento Paralelo Básico
INF-0617 - Big Data

Rafael Fernando Ribeiro
Thiago Gomes Marçal Pereira

Prof. Lucas Wanner

Outubro de 2018

Apresentação do Problema

Para esse primeiro trabalho, foi-nos apresentado um conjunto de livros do Projeto Gutenberg, e o propósito era a contagem de letras/números e definição do carácter mais utilizado em cada livro.

A solução deveria ser realizada através de um processamento serial, e também de um processamento paralelo, para comparação de resultados.

Adicionalmente, resolvemos também através de processamento paralelo com um Pool de processos, respondendo à terceira questão da proposta.

Solução

O projeto foi implementado em Python, onde foi definido um método único para a contagem do caracter mais utilizado em um determinado arquivo e caso desejado, imprimir o nome do arquivo e resultado.

Então, foram definidos outros 3 métodos, para os processamentos.

- Em série (um arquivo após o outro), que faz o uso de um único processo
- Em paralelo, para quantos arquivos estiverem presentes
- Em paralelo com Pool, podendo alterar a quantidade de processos (realizamos inclusive a execução com diferentes, para montagem de um gráfico.

A execução foi feita em 2 ambientes (3 configurações):

- (1) Macbook (4 núcleos 2 Threads) - 16GB RAM SSD
- (2) Windows (i7 4 núcleos 2 Threads) - 16GB RAM SSD
- (3) Windows (i7 4 núcleos 2 Threads) - 16GB RAM Sata

Os resultados obtidos são:

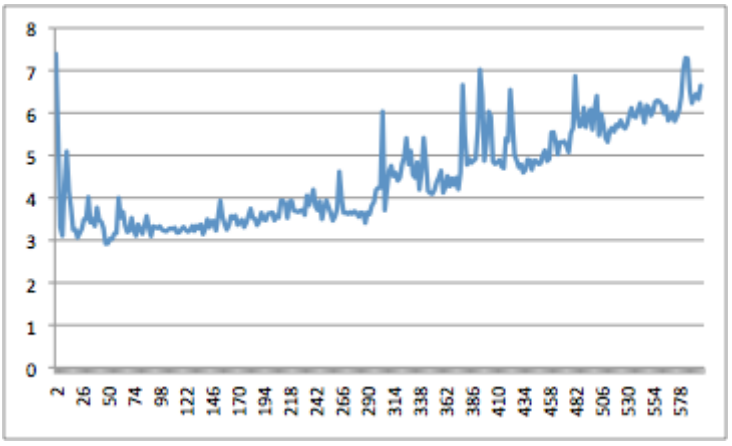
	(1)	(2)	(3)
Proc. Serial	9.81 s	19.52 s	36.46 s
Proc. Paralelo	2.79 s	60.71 s	60.34 s
Proc. Pool	2.62 s	11.98 s	13.43 s

Conclusão

Podemos perceber que o processamento paralelo acaba sendo dependente da quantidade de processos utilizados e como o Sistema Operacional os trata, e qual o custo que isso impacta.

Uma boa maneira de melhorar o processamento paralelo é utilizando Pool de processos, que são reutilizados, e portanto, chegando a um melhor aproveitamento dos núcleos lógicos de processamento.

Como exemplo, podemos ver o gráfico anexo, em que fica claro que após certo ponto, o aumento de processos causa um aumento do tempo de processamento.



Número de Processos vs Tempo de Processamento (Macbook)